English value-added measures

Perry, Thomas

DOI:
10.1002/berj.3247

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Perry, T 2016, 'English value-added measures: examining the limitations of school performance measurement', *British Educational Research Journal*, vol. 42, no. 6, pp. 1056–1080. https://doi.org/10.1002/berj.3247

BERJ British Educational Research Journal

BERA

British Educational Research Journal
Vol. ●●, No. ●●, ●● 2016, pp. ●●–●●

DOI: 10.1002/berj.3247

English Value-Added Measures: Examining the Limitations of School Performance Measurement

Thomas Perry*
University of Birmingham, Birmingham, UK

Value-added 'Progress' measures are to be introduced for all English schools in 2016 as 'headline' measures of school performance. This move comes despite research highlighting high levels of instability in value-added measures and concerns about the omission of contextual variables in the planned measure. This article studies the impact of disregarding contextual factors, the stability of school scores across time and the consistency of value-added performance for different cohorts within schools at a given point in time. The first two analyses replicate and extend previous studies using current data, confirming concerns about intake biases and showing that both secondary and primary level value-added measures exhibit worrying levels of instability. The third analysis goes further by examining whether instability across time is likely to stem from differences between cohorts and whether measures based on a single cohort reflect school performance more generally. Combined, these analyses suggest a general problem of imprecision within value-added estimates and that current policy use of value-added is unjustified. Published school performance measures are likely to be profoundly misleading, in particular for those unfamiliar with the level of uncertainty in the estimates. The article closes by considering whether value-added measures could and should be used by policy-makers as measures of school performance.

Keywords: value-added; accountability; performance measures; school effects; Progress 8

Introduction

This article examines the fitness for purpose of the English value-added (VA) measures of school performance, bringing together results from several analyses. New evidence is presented concerning the consistency of VA scores between different school cohorts at a single point in time. Evidence is also presented regarding the level of observable bias in the current English school VA measure and the stability of this measure over time at both primary and secondary level. This study comes at an important time: VA 'Progress' measures are to be introduced as 'headline' measures of school performance in the 2016 performance tables at Key Stage 2–4 (KS2–4) (DfE, 2014b) and KS1–2 (DfE, 2016a). This move, for the first time, positions VA measures as the key performance indicators for English schools. The analyses in this article, however, show that the existing VA measure (2011–2015) exhibits marked

*School of Education, University of Birmingham, Edgbaston, Birmingham, United Kingdom, B15 2TT. Email: t.w.perry@bham.ac.uk; Twitter: @TWPerry1

biases relating to intake characteristics; that there are serious levels of instability in school-performance scores over time, particularly at primary level; and that a school-performance measure based on a single cohort reveals little about the performance of other cohorts in the school. As a result, the new English 'Progress' VA measures are a severely limited measure of school performance and any high-stakes use is ill-advised.

This article begins by introducing VA and its recent English policy use, including details of the new Progress measures. It then reviews the available literature on stability and bias within primary and secondary VA measures. Following this, the article examines three empirical questions:

1. *To what extent are observable biases present within the existing (2011–2015) English VA measures?* This extends the results of Burgess and Thomson (2013) to estimate school-level biases across a number of contextual factors at both primary and secondary level.
2. *How unstable are the existing VA measures?* This analysis replicates previous studies that examined the stability of the earlier contextualised value added (CVA) measure used in England between 2005 and 2010 (Leckie & Goldstein, 2009; Gorard *et al.*, 2012). The complexity of the CVA model was cited as a factor contributing to its instability and so analysing the simpler VA measure that replaced it assesses whether the situation has improved and the impact of this policy change. The article also extends the analysis to consider the stability of primary-level measures.
3. *To what extent is the VA performance of different cohorts in a school at a given point in time consistent?* The purpose of this final question is twofold: First, it assesses whether the existing school VA measures reflect school performance more generally rather than simply the performance of the cohort on which they are based. Second, it enables more accurate location of the source of the instability in VA models over time. Note that these concerns address the issue of consistency from the perspective of school-level performance measurement. While the underlying causes of any within-school variability are an important research area, with the introduction of the new Progress measures, school-level measurement is the most pressing issue at present. The question of within-school consistency is addressed by looking at the correlation between CVA performances of cohorts within schools in a single year. This analysis uses teacher-assessed levels and a simulated CVA measure for each cohort in a large dataset obtained by the Department for Education (DfE) spanning National Curriculum (NC) years 3 to 9.

Examining bias, stability and consistency concurrently allows links between these to be examined and the overall validity of school VA measures and their limitations to be evaluated. The introduction of the new Progress measures makes up-to-date estimates in these three areas of great value.

The logic of VA

VA promises to enable fair comparisons of school performance despite schools having markedly different pupil intakes. By identifying and controlling for factors that lie beyond schools' control, VA measures are designed to compare schools only on the basis of unexplained residual variation between (statistically) 'like-for-like' pupils.

Remaining differences between school-level examination results are treated as evidence of relatively high or low school effectiveness (Nuttall *et al.*, 1989). A simple approach to this is to compare the performances of a particular group of pupils to the performance of other pupils with the same examination score at an earlier point in time. The mean performance of other pupils with the same earlier score can be taken as an 'expected' or 'estimated' score and used as a baseline against which a pupil's relative performance can be evaluated. A positive or negative VA score is produced from the difference between a pupil's actual score and this statistical expectation. To avoid negative scores being interpreted as negative progress (rather than lower than average progress), the DfE previously added a fixed number (either 100 or 1000) to the VA score (Ray, 2006). This practice is due to cease with the new Progress measures which will remain centred on zero (DfE, 2016b).

VA works on an eliminative logic: by controlling for the effects of non-school factors one can attribute the remaining variation to differences in school effectiveness. Following this logic, more complex statistical models can be used to create performance expectations (i.e., comparisons) taking a large number of measured non-school factors into account in order to isolate more completely the difference in performance attributable to differences in school effectiveness from myriad other influences, making the comparisons as fair as possible.

English policy use of VA

English VA scores were first published in 2002. They use performance data collected for all state school pupils in the English system at the end of a number of 'Key Stages', age 7, 11, 14 and 16 (Key Stages 1 to 4, respectively). Particularly important are the KS2 (aged 11) and the KS4 (aged 16) results as these mark the end of primary- and secondary-level education. Using the VA method, performance at one key stage can be compared to similar pupils with the same scores at an earlier key stage.

Between 2005 and 2010, secondary schools in England were compared using a contextualised value added (CVA) model which, in addition to controlling for prior attainment at earlier key stages, attempted to account for a wide range of non-school factors associated with students' progress (such as eligibility for free school meals). It was an attempt to achieve a fair, unbiased comparison between pupil performances with the data that were available. In contrast, the VA measure that replaced it (2011–2015) ignored all contextual variables, comparing pupil performances only in relation to attainment at previous key stages. This was a political decision made by the government who felt that it was 'wrong in principle' (DfE, 2010, p. 68) to take characteristics other than prior attainment into account.

The new generation of Progress measures

In 2016, new VA measures known as 'Progress' measures are to be published for secondary and primary schools in England. There is due to be a KS2–4 measure and a KS1–2 measure. Although the underlying attainment measures differ, the VA method works in the same way and so only the KS2-4 measure is discussed in this section. The key measure of attainment at the end of KS4 is named Attainment 8. This

measure is created using a series of tariffs giving different values to various qualifications at different grades in the KS4 national examinations (see DfE, 2016b for further details). This section puts issues with the score tariffs to one side and, instead, details how the Progress 8 score is produced.

The KS2–KS4 'Progress 8' measure is calculated by adjusting Attainment 8 scores using pupils' prior attainment scores at KS2. Progress 8 is much simpler than previous measures. Pupil KS4 performances are adjusted using a single variable: an average point score (APS) of English and mathematics performance at KS2. Progress 8 uses the mean APS performance of pupils with the same attainment at KS2 as the baseline for VA comparisons at KS4. Using a single prior attainment variable and ignoring contextual factors eliminates the need to use complex statistical models to compare results across numerous variables. Using simple comparison tables has been found to produce almost identical estimates to those that would be produced by the multi-level regression model used for the 2011–2015 VA measure (Burgess & Thomson, 2013). Moreover, simplicity is likely to be advantageous in some respects: it is more likely than CVA to be stable over time (Dumay *et al.*, 2013) and has the potential to be more widely understood than previous measures, which were considerably more complex (Kelly & Downey, 2010).

Like the measure it replaces, Progress 8 is to ignore contextual factors. This was explicitly stipulated as a requirement for its design (Burgess & Thomson, 2013). Consequently, the measure is knowingly disadvantageous to schools with a disproportionate number of pupils whose characteristics are associated with lower performance such as those classified as being in poverty. This places a limit on the extent to which the measure can be said to be 'unbiased' and 'fair' (Burgess & Thomson, 2013). The implications of bias are particularly concerning since the measure is also planned to be used as a basis for a 'floor standard' of performance, identifying low-performing schools requiring intervention. In their report on the new Progress 8 measure, Burgess and Thomson (2013) show that, due to the exclusion of contextual factors, schools falling below the floor standard are likely to serve localities with high rates of poverty. The combination of this bias with a floor standard of performance will mean that, 'schools in disadvantaged areas may face continuous intervention' (Burgess & Thomson, 2013, p. 17). Progress 8 is also likely to be biased in relation to other contextual factors such as those included in previous CVA models (Evans, 2008). Pupil-level estimates of biases provided by Burgess and Thomson (2013) are consistent with these but school-level estimates are not given. The magnitude of biases in the school results is not clear from pupil-level results alone as it will depend on how high- or low-performing pupil groups are distributed across schools. Also, as Burgess and Thomson's report pertains to the secondary performance measure, no primary-level data were presented. These issues are addressed in the first empirical section in this article, where school-level estimates of bias in recent VA scores at secondary and primary level are presented.

One final noteworthy aspect of the planned Progress 8 measure for future years are the plans to set the expected level of performance in advance. This approach would make use of the associations between KS2 and KS4 performances from previous cohorts (an *ex ante* model) rather than waiting for the actual cohort data to be available before making performance comparisons, as is currently practiced. This would

have the advantage that schools would know the baseline against which their pupils' actual scores will be evaluated in advance; although whether this will be workable or will have any beneficial effect on school or pupil behaviour is yet to be seen. A final decision on *ex ante* models will be made in 2017 (DfE, 2014a).

Biases and errors in VA estimates

This section reviews the most recent evidence about stability and bias, which casts doubt on the capacity of VA models to provide acceptably accurate and fair comparisons between schools. Many of these difficulties have been discussed at length by educational effectiveness researchers over several decades (see, for example, Sammons, 1996; Goldstein, 1997; Teddlie & Reynolds, 2000; Visscher, 2001).

Bias and confounding

As described above, VA measures try to 'level the playing field' when making performance comparisons by controlling for non-school influences. Any remaining differences will bias the measure. The 2007 CVA model included measures of deprivation, in care status, special educational needs status, pupil mobility, gender, age within year, English language status, ethnic group and school average prior attainment (Evans, 2008). These are in line with those consistently found to be associated with performance in school-effectiveness research (Teddlie & Reynolds, 2000). The impact of these on the measure is small relative to that of prior attainment. Prior attainment, as well as reflecting a direct effect of previous performance supporting future learning, acts as a 'black box' that indirectly accounts for a large number of unobserved factors that have brought about the differences in previous performances. Nevertheless, for schools with challenging intakes, the impact of contextual variables on the VA scores can be substantial.

Attempting to eliminate all biases caused by non-school factors makes high demands on the available data (Goldstein, 1997; Coe & Fitz-Gibbon, 1998; Gorard, 2010). Even in high-quality datasets such as the National Pupil Database (NPD), there are large amounts of missing data (Gorard, 2010) and known non-school factors for which data are unavailable (Dearden *et al.*, 2011). Studies into measurement error, such as Ferrão and Goldstein (2009, p. 963), suggest that errors that are randomly distributed across schools and pupils tend not to have a large impact on VA estimates, although they note that this does not account for 'unobserved patterns of measurement error variation across schools' (also see Gorard, 2011 on the assumption that errors are randomly distributed).

Even if the data were comprehensive, complete and error-free, accounting for non-school factors remains problematic. School-effectiveness researchers have long known that correcting for intake differences can also unintentionally remove genuine differences in school performance (Visscher, 2001, p. 207). Associations between non-school factors and performance say little about the underlying cause. Take, for example, the finding that poverty is associated with lower school performance. If this was due to external influences, not taking this into account would disadvantage schools in poorer areas. If, however, the difference reflected poorer standards of

education in poorer areas, it may be inappropriate (and, indeed, 'wrong in principle') to remove this difference. For each available contextual variable, judgement is required as to which of these explanations is most likely and, therefore, the biggest threat to validity. This is an approximate process with no entirely satisfactory solution (Visscher, 2001). The Progress measures avoids these difficulties by simply not attempting to take anything other than prior attainment into account.

Stability over time

Several researchers have looked at the level of stability in the English CVA measure (2005–2010), finding moderate levels of stability (Leckie & Goldstein, 2009; Gorard *et al.*, 2012). The most recent estimates come from Gorard *et al.* (2012), who presented the correlations of school CVA scores 1, 2, 3 and 4 years apart finding correlations ranging from 0.58–0.79, 0.48–0.67, 0.56 and 0.46 respectively. These results show that, even 1 year apart, there is only a moderate correlation in school CVA scores. Gorard *et al.* (2012, p. 7) reached the conclusion that the CVA scores appear to be 'meaningless'. A slightly more positive conclusion is reached by Allen and Burgess (2013) who, looking over several years, found that the use of performance information for school selection nearly doubles the chance of selecting a good school, although they found that 'the best performance information is only slightly more useful in school choice than a school's composition, measured by the average prior attainment of pupils entering the school.' (Allen & Burgess, 2013, p. 186). It may be that the continuity in performance is partly due to enduring non-school factor biases such as pupil composition.

Other research has produced similar results in different areas and for different ages. Primary-level school VA scores are found to be highly unstable. Dumay *et al.* (2013) looked at VA performance of different primary grades across time for 1, 2, 3 and 4 years apart, finding correlations of 0.40–0.53, 0.40–0.43, 0.36–0.40 and 0.29 respectively. The only thing that was stable was that the vast majority of schools had 'indeterminable' effectiveness (Dumay *et al.*, 2013, p. 75). Similarly, research into systems other than England has found very low correlations in performance. Marks (2014, p. 14) estimated year-on-year correlations in VA performance in grades 5 and 9 as ranging from 'from a very low 0.10 to 0.30 for Year 5 and from 0.16 to 0.50 for Year 9'. Instability is a long-standing problem in primary-level VA scores (Tymms & Dean, 2004, p. 5).

There are several possible explanations for the levels of instability in VA measures over time: differences between schools' effectiveness may not be meaningfully large or stable; fluctuations in uncontrolled (perhaps unobserved) non-school factors or measurement error may be substantial relative to differences between schools; or perhaps more complex contextualised models are 'over-correcting' for intake differences, removing some genuine differences in school effectiveness along with the influences of non-school factors (see previous section). Allen and Burgess (2011, p. 253) suggest that 'CVA is unstable because it results from fitting a complex model with many imprecisely measured parameters'. Similarly, Gorard *et al.* (2012) note variation in the model coefficients over time as well as stressing a number of problems with the quality of the underlying data, especially in relation to the measured contextual

variables. Measurement error may also be a problem in the measures of attainment with the combination of error in the prior attainment measure and the main performance measure having the potential to 'propagate', inflating the relative magnitude of the error considerably beyond its original scale (Gorard, 2012).

Modelling choices must take into account the important links between stability, error and bias (i.e., reliability over time and internal validity) (Dumay *et al.*, 2013). Using a greater number of contextual variables is likely to reduce bias but is likely to make the resulting score more unstable and make greater demands on the quality of the data.

Consistency between cohorts

Instability between cohorts is often suggested as a reason for observed levels of instability over time (Dumay *et al.*, 2013; Marks, 2014). Yet, there are very few examples of studies that have been able to present evidence concerning the consistency of relative VA performances across different cohorts (year groups) at the same point in time in the same school (Teddlie & Reynolds, 2000). An early key study identified by Teddlie and Reynolds (2000) is Mandeville and Anderson (1987) who find 'discouragingly small' correlations and characterise consistency between grades 1 through 4 as 'very unstable' (pp. 212 & 203). Subsequent work (Bosker & Scheerens, 1989) suggested more moderate correlations between grade levels but urge caution, suggesting the figures were possibly inflated. As Teddlie and Reynolds (2000, p. 118) noted in 2000, the question '[had] not been adequately researched'. A literature search conducted during this study was unable to find any more recent studies.

Examining the consistency of estimates between cohorts is of great value for two main reasons: First, this addresses the question of whether school-performance measures based on a single cohort reflect performance at the school more generally, and so whether school-level VA measures have meaningful internal consistency. Second, this gives an indication of whether instability (see earlier) stems from fluctuations in school performance over time or from factors at a given point in time giving rise to inconsistency between cohorts such as differences in teacher effectiveness, unobserved differences between cohorts and more general problems of measure validity.

Method

Research questions

This research examines threats to validity in VA measures. Below are the results of three simple analyses – each focusing on a key threat to validity. These can be summarised by 3 research questions:

- Analysis 1 – What is the impact of observable biases on the 2014 VA measures?
- Analysis 2 – How stable are the 2011–2014 VA measures?
- Analysis 3 – How consistent are VA estimates of performance between cohorts in a given school in a given year?

These three questions are all addressed at both primary and secondary level using English data. The first two analyses use readily available school-level data from the NPD containing data for nearly all state schools in England. The first two analyses are based on previous studies, replicating and extending the evidence using a) the 2014 VA model, b) data at both primary and secondary level and c) school-level estimates of biases. Pupil-level data from 2012 and 2013 are also used for some of the supplementary results and robustness checks mentioned.

The third analysis is based on a large dataset obtained by the DfE. It is a rare example of a large English dataset containing performance of pupils between key stages. It contains teacher-assessed performance data for 2008–2010 for NC years 3–9.

Data source 1: School-level National Pupil Database 2011–2014

School-level NPD data for all state schools in England are readily available from the DfE performance tables website (DfE, 2015). This excludes approximately 7% of pupils nationally who attend private schools. The VA measures, underlying performance measures and contextual variables mentioned later all refer to these official data as published in the public performance tables for purposes of parental choice and school accountability. All special schools were excluded from the analysis and schools were matched using establishment numbers to ensure continuity for schools converting to 'academy' status during the recent reforms taking place in England.

Data source 2: 'Making Good Progress' data 2008–2010

For the third analysis, data were obtained from a study conducted by the DfE known as 'Making Good Progress' (MGP) which looked at how pupils progressed during Key Stages 2 and 3 (DfE, 2011). The MGP dataset is large, with data for 148,135 pupils spanning 342 schools, 10 local authorities, 6 consecutive school year groups (UK years 3–9) across 3 years (2008–2010). There were 100,000 pupils in 2007/2008 with pupils being fairly evenly spread across years 3 to 9 (age 8 to 14). This overall number dropped to just over 70,000 by the third year, again spread fairly evenly across the age range. The MGP report compares the achieved sample across a range of pupil background variables with national data for these year groups, finding it to be 'broadly representative' of pupils in years 3 to 9 nationally (DfE, 2011, p. 6). Further details of the MGP dataset can be found in the DfE report (DfE, 2011), including more details on the sample (p. 5), teacher-assessed performance (p. 7 and see later), mean attainment scores and their distribution (p. 19) and factors associated with differences in performance (p. 24).

The analyses using the MGP data in this article uses the teacher-assessed mathematics outcome data. Scores are recorded in NC sub-levels, with each sub-level representing about 8 months of typical progress. Teachers used evidence of pupils' work, their professional knowledge and the results of classroom tests to assess pupils' attainment level against best-fit NC level descriptors. These are the highest quality performance measure available for all years in the MGP data. Examination-based national Key Stage results are available for NC years 6 and 9, but the analyses using the MGP data in this article require scores for the NC years in between these end of Key Stage years.

Use of teacher-assessed NC scores has some notable limitations: NC levels are designed to be a single scale that tracks attainment from age 5 to age 14. It is questionable, however, whether the scale and the teacher-assessment is consistent across teachers across the full age range. There is evidence to suggest that teacher-assessed levels can be unreliable in some circumstances yet this can be improved by moderation procedures and well-designed assessment criteria (Harlen, 2005). The evidence base on both the reliability of teacher assessments and the effectiveness of moderation in improving it is, however, considerably lacking at present (Johnson, 2013). The teacher-assessed mathematics scores showed relatively high consistency with scores obtained through external examinations and little tendency to be systematically biased in a particular direction. Agreement between teacher-assessed levels and examination-based levels ranged from 64% to 89% in KS2 mathematics (DfE, 2011, p. 41). Some of the discrepancy stems from differences in timing between the two measures, with teachers scores being lower than the examined results due to being recorded some time earlier (DfE, 2011). Results for the second and third study years are likely to be more robust as the correspondence between the teacher-assessed levels and the examination levels increased over the time period from 2008 to 2010 due to moderation activities, with the quality of the teacher assessments improving as the 'processes bedded'(DfE, 2011, p. 7). The sample in the third year was slightly reduced, especially at secondary level where the school numbers were much lower.

Inspection of the data revealed concerns regarding the KS1 and KS2 scores collected in national examinations that were matched with the teacher-assessed data in the MGP dataset. The KS1 data were particularly problematic, showing a marked ceiling effect (affecting 8% of pupils) and a spike at the expected score of 15 (16% of pupils). These findings are problematic for use of KS1 data as a baseline in the KS1–2 measure but do not appear to have greatly affected the estimates in Analysis 3, on consistency: When Analysis 3 was repeated without pupils at the ceiling and expected scores, consistency modestly decreased suggesting that consistency estimates are slightly inflated by problems with the baseline data. Further details of the MGP data used in the analysis are given in Table A1, Appendix A.

Results from Analysis 1 - Observable biases

Primary and secondary school VA measures in 2014 were regressed on a number of contextual factors to assess the extent of systematic biases in the measures. The following school-level variables were included: special educational needs (SEN) (%), English as an additional language (EAL) (%), disadvantage as measured by free school meals eligibility and child looked after status (%), number of pupils in cohort, average cohort prior attainment (at KS1/KS2), percentage of eligible pupils who are male and coverage (inclusion in the measure) as a percentage of eligible pupils. Collectively, these variables accounted for 11% (R^2) of variance in the primary school VA scores and 35% (R^2) of variance in the secondary school scores.

These observable biases are shown on a series of graphs (Figures 1 and 2). Each graph puts school value-added on the y-axis against a school context factor in a bivariate comparison. Each data point in Figure 1 is a primary school in England. There is a linear trend line on each scatter plot showing the systematic relationship between

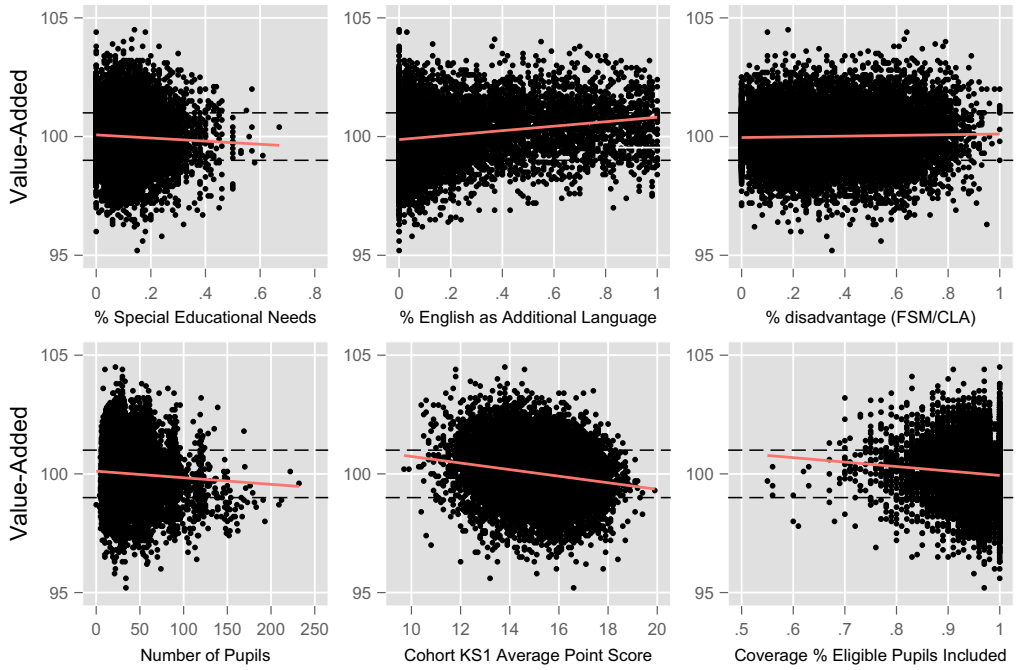

Figure 1. 2014 School KS1-2 Value-Added Scores against selected Contextual Variables at School-Level

VA performance and the factor in question. Also, each plot has two horizontal reference lines set at 1 NC average point score *per pupil* above and below expected performance, this equates to about 4 months' extra/lower progress since the previous key stage, 4 years earlier. These plots have been checked against the output from the regression analyses (see Appendix B) to ensure that bivariate comparisons are not misleading due to countervailing effects between factors.

These scatter plots show a number of small to moderate biases. The most substantial biases relate to the number of pupils with EAL, where schools with higher rates of EAL are linked with higher performance, and the cohort's average key stage 1 score (age 7), where higher performing cohorts at KS1 are less likely to get high KS1–KS2 VA. It appears that schools that maximise KS1 performance find it more difficult to make KS1–2 VA.

One noteworthy difference between the bivariate graphical analysis and the multivariate analysis at primary level is the lack of cohort-level association between disadvantage and VA. The multiple regression analysis (Appendix B, Table B2) found that disadvantage was associated with a decline in school VA performance of about 0.1 NC points for each 10% of pupils on Free School Meals (FSM). This suggests that countervailing effects (such as a tendency for schools with high EAL rates to also have high FSM rates) are at play and so the multivariate results are likely to be more accurate as an estimate of effect for this variable. The negative association between FSM and VA performance at school level is in line with the recent study by Strand (2016) who found a negative cohort-level association between FSM and performance

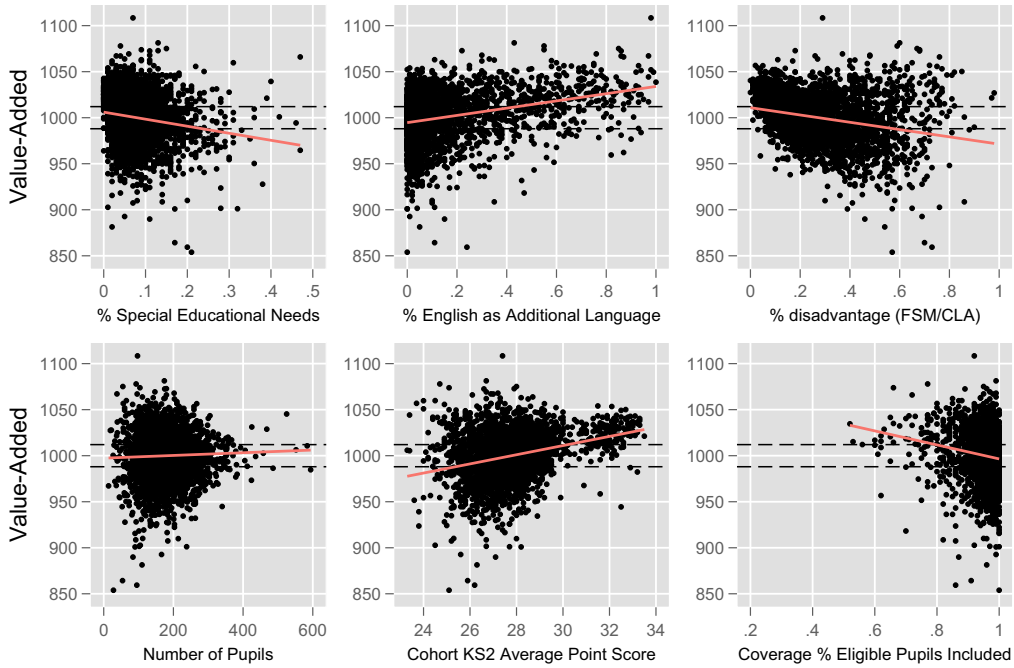

Figure 2. 2014 School KS2-4 Value-Added Scores against selected Contextual Variables at School-Level

even after controlling for individual pupil FSM status. Pupil-level analyses in the present study (see later) found that FSM eligibility was associated with lower performance of about 0.33–0.46 NC points, varying according to which other contextual variables were accounted for in the CVA model.

There are greater rates of observable bias at secondary level with a number of contextual variables showing large systematic relationships with school VA scores. The scatter plots in Figure 2 are the equivalent of those in Figure 1 but at secondary level. The horizontal reference lines are set at 12 points on the 'Best 8' attainment measure above and below the statistical expectations. That equates to two GCSE (or equivalent) grades on average per pupil above or below those expected across the pupils' best eight examination grades. A pupil expected to achieve eight GCSE C grades at a school on the upper reference line, for example, would instead receive two B grades and six C grades.

Figure 2 shows a number of moderate to large systematic relationships between school VA and intake characteristics. The key negative bias relates to the proportion of disadvantaged pupils (as indicated by free school meals or looked after status). SEN status appears important in the graphical analysis but has a far smaller effect in the regression analysis which accounts for countervailing effects. Large positive biases are apparent for schools with a high proportion of pupils with EAL. There is a sizable negative relationship between coverage and VA performance; regression analysis upholds the size of this effect although suggests it is highly inconsistent. As these pupils will not be missing at random (Gorard *et al.*, 2012), low coverage may appreciably bias scores for affected schools. Cohort KS2 average point scores exhibit a

positive relationship, largely driven by the cluster of selective 'grammar' schools. This is in line with previous research (Leckie & Goldstein, 2009) and recent analysis undertaken at the independent research centre, Education Datalab (Allen, 2015). This particular problem highlights the 'fragility' of causal inferences based on VA measures (Marsh *et al.*, 2011, p. 282) introduced earlier. It might be that grammar schools are more effective; this may be a compositional effect stemming from the grouping of higher ability pupils or it could be a 'phantom' effect arising from poor quality data or unobserved non-school factor bias (Harker & Tymms, 2004, Televantou *et al.*, 2015). As Harker and Tymms (2004, p. 195) note, 'the really worrying thing is that the researcher can never be sure about what has been found'. A small positive effect (just under four points) was also found for schools with single-sex intakes in further regression analyses (not shown), even after controlling for grammar school status and gender. This suggests either advantages of single sex schooling (for both genders) or unobserved factors are present.

As a final estimate of the impact of disregarding contextual variables, a CVA replica measure was created using pupil-level NPD data, which adjusted for the contextual variables used in the regression models above. The difference between the official VA score and the CVA replica scores was calculated to indicate the extent to which school scores would change if these contextual variables were accounted for. At KS2–4 the difference ranged from -33.7 to 33.5, with a standard deviation of 7.5. At KS1–2, the difference ranged from -1.5 to 1.6, with a standard deviation of 0.34. These figures show that the policy decision not to account for contextual variables has large effects on the scores of many schools.

Results from Analysis 2 – Stability of VA over time

The second analysis in this article replicates previous studies of the stability of the English CVA scores using the 2011–2014 VA model and extends the analysis to consider primary-level scores. The official school-level VA scores for all state schools

Table 1. Pairwise correlations over time in value-added and unadjusted performance

Primary level	School value-added			Unadjusted performance (APS)		
	1 year earlier	2 years earlier	3 years earlier	1 year earlier	2 years earlier	3 years earlier
2014	0.61	0.46	0.35	0.66	0.61	0.56
2013	0.60	0.45		0.66	0.60	
2012	0.59			0.66		
Secondary level	**School value-added**			**Capped GCSE point score (unadjusted)**		
	1 year earlier	2 years earlier	3 years earlier	1 year earlier	2 years earlier	3 years earlier
2014	0.56	0.49	0.44	0.70	0.62	0.60
2013	0.79	0.68		0.94	0.93	
2012	0.79			0.96		

from the NPD were used to estimate correlations between raw and VA performance for results lagged by 1, 2 and 3 years. These are shown in Table 1.

At primary level, unadjusted performance correlations over time have moderate stability. The VA correlation between the current and previous year is roughly similar to the unadjusted correlation. However, the VA correlations fall quite sharply when comparing VA scores 2 and 3 years apart. With a correlation of 0.35, primary school VA performance is weakly related to performance 3 years earlier. A correlation of 0.35 means that only about 12% of the variance is common to both years.

As at primary level, secondary raw-score correlations are more stable over time than the VA correlations. Both raw and VA correlations fell sharply in 2014, presumably reflecting examination reforms that reduced the number of qualifications included in the attainment measures. It is likely that the 2012–2013 results are more representative of normal conditions. These VA correlations are moderate and are on the high end of the correlations found by Gorard *et al.* (2012) concerning the CVA measure of between 0.58–0.79 and 0.48–0.67 for 1 and 2 years apart, respectively, suggesting that the stability of secondary VA is higher using a VA model than with the CVA model. Given the observable biases revealed in the previous analysis, however, slightly improved stability may not be a good thing. Relatively more stable contextual influences remaining in the data will have shifted the unstable CVA scores towards the more stable but more biased raw scores.

Results from Analysis 3 – Consistency in cohort performances

This section analyses consistency in performance for cohorts within schools at a point in time. As this required performance scores for pupils between KS examination years, the MGP data were used (see earlier). A fairly simple CVA model was used to estimate relative progress of each cohort since the previous KS examination. This specification offers a reasonable compromise between bias and potential instability (see earlier). A VA model without the contextual variables was also produced but, as this gave highly similar results, these are not reported in detail. The following model was estimated for the performance of *i* pupils in *j* schools, using data for one cohort at a time:

$$1)\ \text{Performance}_{ij} = \beta_{0j} + \beta_1 \text{KS1APS}_{ij} + \beta_2 \text{KS1APS}_{ij}^2 + \beta_3 \text{Gender}_{ij} + \beta_4 \text{FSM}_{ij} + \epsilon_{ij}$$

$$2)\ \beta_{0j} = \gamma_{00} + u_{0j}$$

Where performance is regressed on the previous key stage's average point score (either KS1 or KS2) and this squared as well as two binary variables, one for pupil gender and free school meals status, an indicator of poverty.

β_{0j} is calculated in a multi-level regression model separating an overall school intercept (γ_{00}) from the differential school intercepts around this point (u_{0j}), giving the estimated school effects for each cohort.

The correlations below are based on the results of 251–271 schools in 2008 and 2009 and 207–226 schools in 2010. Secondary-level correlations are based on the results of 67–71 schools in 2008 and 2009 and 48–51 schools in 2010.

Table 2 shows the correlation between the CVA performance of a NC year group at a school and the CVA performance of other cohorts 1, 2 and 3 years lower (where available) in the same school at the same time. For example, the correlation between the CVA performance of Year 5 cohorts in 2009 and Year 4 cohorts (1 year lower) in the same school is 0.52. The correlation between 2010 year 6 cohorts and year 4 cohorts (2 years lower) at the same school is 0.24.

Table 2 shows correlations at primary level are generally very low to moderately low. This means that knowing the CVA performance of Year 6 in a primary school reveals very little about the performance of Year 3 or 4 in the same school. Even the correlation of consecutive years' performance of around 0.5 is not high (Figure 3).

There are a fewer schools at secondary level. There are also outliers in the data. Given the moderation activities over the 3 calendar years of the study and the drop in school numbers in the final year, the 2009 results are likely to be the most robust estimates, suggesting correlations of about 0.7 and about 0.45 for cohorts 1 and 2 years apart, respectively.

Interestingly, the results show that the results of consecutive NC year groups are more similar than those 2 or even 3 years apart. If this instability were the result of random fluctuations in cohort characteristics or random measurement error, we

Table 2. Pairwise correlations in CVA scores for different cohorts in the same school at the same time for three study years (2008–2010)

Primary level Study year		1 NC year lower	2 NC years lower	3 NC years lower
2008	NC Year 6	0.28	0.13	0.13
2009		0.49	0.27	0.19
2010		0.50	0.24	0.31
2008	NC Year 5	0.46	0.29	
2009		0.52	0.39	
2010		0.59	0.40	
2008	NC Year 4	0.44		
2009		0.48		
2010		0.51		
Column mean		0.47	0.29	0.21

Secondary level Study year		1 NC year lower	2 NC years lower
2008	NC Year 9	0.56	0.52
2009		0.74	0.45
2010		0.38	0.37
2008	NC Year 8	0.58	
2009		0.70	
2010		0.67	
Column mean		0.61	0.44

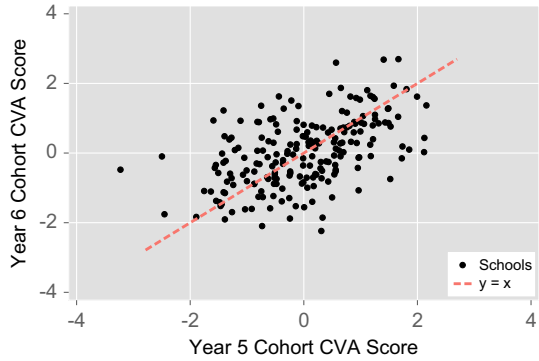

Figure 3. School CVA performance for NC Year 5 and Year 6 in 2010

would expect the correlations in performance between cohorts to be similar irrespective of how many years they are apart. There appears to be some systematic variation, although it is difficult to know its source. It may reflect the likelihood of cohorts separated by a small number of years receiving the same quality of education. As year groups are further separated, they are less likely to have had common inputs (in terms of teachers and curriculum) at various stages in their education. Educational-effectiveness research emphasises the importance of teachers and classrooms rather than schools (Muijs *et al.*, 2014). Given the level of instability in teacher-level VA estimates (Amrein-Beardsley, 2014), however, it is maybe more accurate to understand this explanation in terms of various factors (pupils, teachers, curriculum and myriad complex cultural and pedagogical factors) coming together to create a more or less favourable learning environment for a period of time, which consecutive year groups are more likely to share. While this has intuitive appeal, caution is required: the systematic pattern could equally reflect changes in school intake or in the measures of attainment. As discussed earlier, causal inferences based on such correlational evidence are problematic (Coe & Fitz-Gibbon, 1998; Marsh *et al.*, 2011).

Despite these suggestions of systematic effects, the results give a bleak picture of consistency in cohort performances within schools. This result was examined further in two ways: First, the consistency of raw performance was calculated to see whether inconsistency stemmed from a problem with the teacher-assessed levels. Levels of stability in the teacher-assessed scores were very similar to those based on the examination-assessed GCSE and KS2 scores given earlier in Table 1 (see Table C1 in Appendix C1). Educational-effectiveness research has shown that raw scores tend to be more stable over time than VA measures (Gray *et al.*, 2001; Luyten *et al.*, 2005; Dumay *et al.*, 2013). The present results show that this is also the case for consistency: adjusting for non-school factors such as prior attainment leaves a residual which is comparatively more unstable and inconsistent than the original score.

The second follow-up analysis followed the performance of particular cohorts over time to examine their stability (recall that the MGP data ran from 2008 to 2010). The results are given in Table 3.

The correlations in Table 3 show moderate stability overall, with correlations ranging from 0.43 to 0.73. These figures along with the results from the first follow-up

Table 3. Pairwise correlations in CVA scores between a given cohort's (B–H) performance over study years (2008–2010)

Cohort letter	Year	NC year	1 year earlier	2 years earlier	No. of cohorts	Mean pupils per cohort
Primary level						
B[1]	2010	4	0.65		225	53.8
C	2010	5	0.66	0.52	226	57.7
	2009	4	0.61		272	54.6
D	2010	6	0.57	0.44	212	59.9
	2009	5	0.73		272	56.6
E	2009	6	0.55		260	56.3
Secondary level						
F	2010	8	0.62		52	225.5
G	2010	9	0.69	0.62	49	233.7
	2009	8	0.46		70	223.0
H	2009	9	0.43		68	222.2

[1]The cohorts from the MGP data were labelled from A and I, from youngest to oldest. Cohorts A and I were only in the sample for 1 year so stability cannot be calculated.

analysis suggest that instability and inconsistency is a more general problem in VA measures.

In sum, teacher assessment may have contributed to rates of inconsistency, however this result should be considered alongside the low levels of stability found across years in VA measures, discussed earlier. The official VA measures analysed earlier are largely based on standardised examinations and so inconsistent teacher assessment cannot be the source of the instability in that case. Also note that mathematics scores were used in an attempt to ensure measure reliability, recent research suggests that the consistency of English scores will be considerably lower (Strand, 2016). Further research may wish to replicate these results using more robust tests to estimate the extent to which teacher assessment contributes to inconsistency between cohorts, especially as teacher assessments are used extensively for KS1, KS2 and KS3 assessment as well as being used in day-to-day decisions about performance, curriculum and even teacher pay in English schools.

Discussion and conclusions

This article has presented a clear, up-to-date analysis of the levels of bias and instability in VA measures. The results have huge implications. The analysis of consistency between cohorts is a rare example of a study that has examined inconsistency of cohort performance at a single point in time. It has shown that performance measures based on a single cohort poorly reflect performance in schools more generally. Moreover, a number of substantial biases and low rates of stability have been found in both primary- and secondary-level VA scores. Given the importance of VA measures in educational accountability in England and the introduction of the Progress 8 measure as a headline measure of school performance, these are concerning results.

Observable biases

The 2011–2015 VA measure and the future Progress 8 measure are designed to take only prior attainment into account. Results presented here have confirmed Burgess and Thomson's (2013) concern that this decision will introduce a number of biases into the measure. The biases identified in this study are sufficiently large to undermine the measure's credibility as a fair 'like-for-like' comparison and present serious systematic (dis)advantages for schools relating to pupil intake and school context. This situation is hard to justify given that variables used are readily available in the NPD and could be taken into account, as in the CVA measure. The politicisation of this methodological issue appears to stem from several sources: first, a confusion between statistical and pedagogical expectations; second, the view that it is the standards and expectations set by policy-makers and regulatory bodies that are the key 'driver' of performance; and, third, the political value of the policy-change as a 'gesture' to signal the government's ethos. The real choice in taking contextual variables into account is whether one wishes to make the measure identify and reward schools that have been able to overcome difficult circumstances or whether we wish to punish schools who are unable to do so entirely: English policy makers have chosen the punitive option. It is worth making a clear distinction between statistical and pedagogical expectations. The latter are cultural and can reflect any level of aspiration, whereas the former merely reflects the status quo and, crucially, changes as the situation does. Adjusting expectations according to prior attainment but not other contextual factors is to misunderstand the correlational nature of the exercise. Why do we not also consider it 'wrong in principle' that we have lower expectations of pupils who have performed poorly in earlier key stages, especially given that these differences are strongly related to social class, gender and ethnicity in the first instance? But, similarly to the principle about ignoring contextual factors, ignoring prior attainment would make little sense if we were genuinely interested in making fair and informed judgements about school performances in their given context.

Instability and inconsistency

The present article has shown that recent VA measures have similar but slightly lower levels of stability compared to previous CVA measures. Primary-level measures have been found to be highly volatile. The level of stability at primary level is in line with estimates from recent research by Strand (2016), who found that stability was particularly low in English. This suggests that the average point scores used in the second analysis in this article are comprised of maths scores that are moderately stable and English scores with much lower stability. Strand's results also show a decline in correlations over time as well as raising further questions about consistency in school performance across subjects (also see Telhaj *et al.*, 2009); inconsistency across subject is an issue that is not addressed here but feeds into an overall picture of low to moderate stability and consistency across time, cohorts and subject areas.

When the problem that VA is trying to solve is looked at in perspective, the degree of imprecision that these results suggest is perhaps not surprising. Decades of educational-effectiveness research has confirmed that the vast majority of

differences in performance between pupils are within rather than between schools (Teddlie & Reynolds, 2000). Only around 5–10% of the differences between pupil performances are associated with school membership (Reynolds, 2008; Wiliam, 2010). This well-established and highly consistent finding is easily lost in the political rhetoric around standards. It is important to stress that the evidence is clear that all schools have a massive impact on children's education (Luyten, 2006). VA scores, however, try to isolate comparatively small *relative* effects from the myriad and complex combination of pupil factors affecting performance (Teddlie & Reynolds, 2000; Reynolds, 2008).

VA scores appear to be comprised of a considerable amount of 'noise' relative to the 'signal', casting serious doubt on the viability of VA measures for high-stakes use. The evidence that has been presented and reviewed suggests that VA scores are comprised of a) differences in school effectiveness, b) unmeasured individual or cohort-level differences and c) measurement error within the underlying measures of performance. The results suggest that the latter two components are substantial and that assuming otherwise risks seriously misplaced inferences. The evidence suggests that there is a general level of imprecision, error and bias and this means that, at best, VA scores are only a rough approximation of performance. Moreover, threats to validity such as omitted variable bias are likely to affect individual school scores unevenly and so, for a portion of schools, VA scores will be grossly inaccurate. This suggests that current policy use of VA and the publication of VA scores in school league tables as a measure of school performance is highly misleading.

Communicating uncertainty

There has been little to no attempt to communicate any of the issues discussed above in published performance tables or in guidance documents issued to school leaders such as those cited earlier. The only suggestion that the VA measure may be less than entirely robust is the inclusion of confidence intervals in the performance tables but even this is misleading: Government-published guidance falsely describes confidence intervals as 'the range of scores within which each school's underlying performance can be confidently said to lie' (DfE, 2014a, p. 9). Yet, confidence intervals do nothing to address systematic biases stemming from omitted non-school factors, non-random measurement error, model specification errors or any of the problems with inference and data quality that have been discussed. Moreover, the use of confidence intervals to convey uncertainty is inappropriate and possibly even harmful in this context (Gorard, 2015). A danger is that the limited test provided by confidence intervals will distract from the numerous non-technical difficulties discussed and lead to misplaced 'confidence' in the results. It might be that confidence intervals, far from urging greater caution and awareness of uncertainty are actually performing a rhetorical function of putting 'statistically significant' results beyond professional dispute or discussion.

Valid uses of VA

The most readily defensible response to the evidence presented here and in previous studies is to discontinue any substantial use of VA until it is shown to reflect meaningfully the differences in school performance (Gorard *et al.*, 2012). That is, school VA scores must be shown to be reasonably stable across time; show consistency across outcomes, cohorts and pupil groups; and capture appreciable differences between the performances of schools.

This article ends, however, by pushing back somewhat against this by discussing uses of VA that may be justifiable and seeking to outline some constructive lines of enquiry. A key basis for this optimistic position is that VA scores have been found to be highly consistent with rates of progress in longitudinal data and, therefore, do capture real differences in pupil progress (Perry, 2015), at least to the extent that underlying measures of performance can reliably measure pupil attainment. The problem seems to lie more in the assumption that differences in VA can be causally attributed to school performance rather than merely capturing *unexplained* differences in performance. In other words, VA raises questions rather than provides answers (Demie, 2013). Can anything be done to separate school effects from error and confounding factors? There are several possibilities.

First, it might be that creating measures averaging scores over several years could smooth over some of the volatility, as was suggested in the initial reports looking into VA (Fitz-Gibbon, 1997) and is currently practised in Wales. Looking at results over multiple years is already considered to be essential practice in school-effectiveness research:

> On the basis of existing research it is apparent that estimates of schools' effectiveness based on one or two measures of students' educational outcomes in a single year are of limited value. Ideally, data for several years (three being the minimum to identify any trends over time) and several outcomes are needed. (Teddlie & Reynolds, 2000, p. 126)

One must be careful, however, not to mistake stability for validity (Gorard *et al.*, 2012). Even if meaningfully stable estimates could be produced using averages across time, this still would not mean that differences in performance are necessarily causally attributable to schools. The opposite approach would be to reject an overall school VA score and seek to convey more effectively the complexity of the data. This complexity renders crude accountability judgements between overall school performance of little value. Much information is lost when presenting VA outcomes as group averages, even when broken down by attainment or FSM status, where mean differences between groups tend to be small (Strand, 2016). Despite the apparent enthusiasm for doing so, there is no technical requirement to summarise the output of VA estimates for groups or cohorts using mean scores (OECD, 2008). One simple alternative is to create pupil-level VA scores and present the proportion of pupils scoring in a number of performance bands. This has several advantages: First, all pupils matter to a greater degree than in an overall mean score. Even when using attainment or FSM group means, a good school can afford to 'let down' small numbers of pupils for this to be masked in the average scores (Wilson & Piebalga, 2008). Second, this is simple to interpret and

allows the bands to be chosen to communicate pupils' scores in relation to expectations: what is known about measurement error and the substantive significance of the differences on the outcome scale can be used to set appropriate thresholds for what constitutes being above and below expectations. Third, this approach clearly communicates the extent of within-school variation. Conveying complexity while keeping the data comprehensible has been an ongoing problem with the presentation of VA (Kelly & Downey, 2010; Allen & Burgess, 2011). Fourth, this prevents the need for confidence intervals, which are inappropriate in this context, create technical barriers to understanding and are highly misleading (see earlier). Fifth, these tables can be extended to consider any group of pupils and also different outcomes.

Finally, it is useful to consider the users as well as the usage of VA measures. Policy-makers are at least partially aware of the shortcomings of the measure: Evans (2008, p. 21) points out that VA should be 'used as a basis for discussion in school improvement and inspections, rather than directly driving any rewards or sanctions' and observing the limitation of using any single measure of school performance as a basis for judging school effectiveness. It is possible that, with appropriate information, inspectors, researchers and professionals are able to use VA scores beneficially by skilfully balancing the scores against other pieces of evidence. It is not reasonable, however, to suggest that the general public will be able to reliably draw warranted conclusions about the effect of schools from performance data in league tables in this way. VA scores may be better placed as a monitoring tool (Foley & Goldstein, 2012) for professional use rather than a 'free-standing' evaluative technology for public consumption.

In summary, VA measures fall far short of producing fair and robust measures of school performance, and policy decisions have most likely introduced further biases into the measure. Even in the best school VA measure, a sizable proportion, if not the majority, of observed differences between scores are likely to reflect measurement error and non-school differences rather than genuine differences in performance. If VA is to be used further, it is vital that those drawing inferences from VA evidence understand these issues and take steps to avoid erroneous inferences. The publication of VA scores in public league tables in their current form or the use of VA scores as a key basis for high-stakes decisions is simply not justified by the evidence presented here and elsewhere. With the new Progress measures set to become the headline measures of school performance, the stakes have never been higher.

Acknowledgements

The author would like to thank Peter Davies, Stephen Gorard and the anonymous reviewers for their helpful comments and suggestions. Thanks also go to the ESRC for funding the author's doctoral research programme from which this article arises.

References

Allen, R. (2015) *We cannot compare the effectiveness of schools with different types of intakes*. Blog: Education Datalab. Available online at: http://www.educationdatalab.org.uk/Blog/May-2015/We-cannot-compare-the-effectiveness-of-schools-wit.aspx#.VaAOkq5Viko (accessed 10 July 2015).

Allen, R. & Burgess, S. (2011) Can school league tables help parents choose schools?, *Fiscal Studies*, 32(2), 245–261.

Allen, R. & Burgess, S. (2013) Evaluating the provision of school performance information for school choice, *Economics of Education Review*, 34, 175–190.

Amrein-Beardsley, A. (2014) *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability* (London, Routledge).

Bosker, R.J. & Scheerens, J. (1989) Issues in the interpretation of the results of school effectiveness research, *International Journal of Educational Research*, 13(7), 741–751.

Burgess, S. & Thomson, D. (2013) *Key Stage 4 accountability: Progress measure and intervention trigger*. available online at: http://www.bristol.ac.uk/cubec/portal/ (accessed 13 January 2016).

Coe, R. & Fitz-Gibbon, C.T. (1998) School effectiveness research: Criticisms and recommendations, *Oxford Review of Education*, 24(4), 421–438.

Dearden, L., Miranda, A. & Rabe-Hesketh, S. (2011) Measuring school value added with administrative data: The problem of missing variables, *Fiscal Studies*, 32(2), 263–278.

Demie, F. (2013) *Using data to raise achievement: Good practice in schools* (London, Lambeth Council).

DfE (2010) *The importance of teaching – White paper* (London, Her Majesty's Stationery Office).

DfE (2011) *How do pupils progress during Key Stages 2 and 3? Research report DFE-RR096* (London, DfE). Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/182413/DFE-RR096.pdf (accessed 6 June 2014).

DfE (2014a) *Progress 8 measure in 2016* (London, Department for Education). Available online at: https://www.gov.uk/government/organisations/department-for-education (accessed 13 January 2016).

DfE (2014b) *Progress 8 school performance measure* (London, Department for Education). Available online at: https://www.gov.uk/government/organisations/department-for-education (accessed 13 November 2016).

DfE (2015) *School performance tables*. Available online at: http://www.education.gov.uk/schools/performance/ (accessed: 26 March 2015).

DfE (2016a) *Primary school accountability* (London, Department for Education). Available online at: https://www.gov.uk/government/publications/primary-school-accountability (accessed 2 March 2016).

DfE (2016b) *Progress 8 measure in 2016, 2017, and 2018* (London, Department for Education). Available online at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/497937/Progress-8-school-performance-measure.pdf (accessed 12 January 2016).

Dumay, X., Coe, R. & Anumendem, D.N. (2013) Stability over time of different methods of estimating school performance, *School Effectiveness and School Improvement*, 25(1), 64–82.

Evans, H. (2008) *Value-added in English schools* (London, Department for Children, Schools, Families). Available online at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.9363&rep=rep1&type=pdf (accessed 3 February 2014).

Ferrão, M.E. & Goldstein, H. (2009) Adjusting for measurement error in the value added model: Evidence from Portugal, *Quality & Quantity*, 43(6), 951–963.

Fitz-Gibbon, C.T. (1997) *The value added national project: Final report: Feasibility studies for a national system of value-added indicators* (London, School Curriculum and Assessment Authority).

Foley, B. & Goldstein, H. (2012) *Measuring success: League tables in the public sector* (London, British Academy).

Goldstein, H. (1997) Methods in school effectiveness research, *School Effectiveness and School Improvement*, 8(4), 369–395.

Gorard, S. (2010) Serious doubts about school effectiveness, *British Educational Research Journal*, 36(5), 745–766.

Gorard, S. (2011) *Comments on 'The value of educational effectiveness research'*. BERA Conference: BERA. Available online at: http://rab.bham.ac.uk/pubs.asp?id=049c1113-9b7a-4978-b370-98cdc55d3e3a (accessed 12 January 2016).

Gorard, S. (2012) The propagation of errors in experimental data analysis: A comparison of pre- and post-test designs, *International Journal of Research & Method in Education*, 36(4), 1–14.

Gorard, S. (2015) Rethinking 'quantitative' methods and the development of new researchers, *Review of Education*, 3(1), 72–96.

Gorard, S., Hordosy, R. & Siddiqui, N. (2012) How unstable are 'school effects' assessed by a value-added technique?, *International Education Studies*, 6(1), 1–9.

Gray, J., Goldstein, H. & Thomas, S. (2001) Predicting the future: The role of past performance in determining trends in institutional effectiveness at A level, *British Educational Research Journal*, 27(4), 391–405.

Harker, R. & Tymms, P. (2004) The effects of student composition on school outcomes, *School Effectiveness and School Improvement*, 15(2), 177–199.

Harlen, W. (2005) Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes, *Research Papers in Education*, 20(3), 245–270.

Johnson, S. (2013) On the reliability of high-stakes teacher assessment, *Research Papers in Education*, 28(1), 91–105.

Kelly, A. & Downey, C. (2010) Value-added measures for schools in England: Looking inside the 'black box' of complex metrics, *Educational Assessment, Evaluation and Accountability*, 22(3), 181–198.

Leckie, G. & Goldstein, H. (2009) The limitations of using school league tables to inform school choice, *Journal of the Royal Statistical Society*, 172, 835–851.

Luyten, H. (2006) An empirical assessment of the absolute effect of schooling: Regression-discontinuity applied to TIMSS-95, *Oxford Review of Education*, 32(3), 397–429.

Luyten, H., Visscher, A. & Witziers, B. (2005) School effectiveness research: From a review of the criticism to recommendations for further development, *School Effectiveness and School Improvement*, 16(3), 249–279.

Mandeville, G.K. & Anderson, L.W. (1987) The stability of school effectiveness indices across grade levels and subject areas, *Journal of Educational Measurement*, 24(3), 203–216.

Marks, G.N. (2014) The size, stability, and consistency of school effects: Evidence from Victoria, *School Effectiveness and School Improvement*, 26(3), 397–414.

Marsh, H.W., Nagengast, B., Fletcher, J. & Televantou, I. (2011) Assessing educational effectiveness: Policy implications from diverse areas of research, *Fiscal Studies*, 32(2), 279–295.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H. & Earl, L. (2014) State of the art – teacher effectiveness and professional learning, *School Effectiveness and School Improvement*, 25(2), 231–256.

Nuttall, D.L., Goldstein, H., Prosser, R. & Rasbash, J. (1989) Differential school effectiveness, *International Journal of Educational Research*, 13(7), 769–776.

OECD (2008) *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools* (Paris, Organisation for Economic Cooperation and Development).

Perry, T. (2015) Measures of school effectiveness: A test of inter-method reliability, paper presented at the *16th Biennial EARLI Conference 2015*, Limassol, Cyprus.

Ray, A. (2006) School value added measures in England,*A paper for the OECD Project on the Development of Value-Added Models in Education Systems* (London, DfES).

Reynolds, D. (2008) *Schools learning from their best: The within school variation (WSV) project* (Nottingham, National College for School Leadership).

Sammons, P. (1996) Complexities in the judgement of school effectiveness, *Educational Research and Evaluation*, 2(2), 113–149.

Strand, S. (2016) Do some schools narrow the gap? Differential school effectiveness revisited, *Review of Education*, 4(2), 107–144.

Teddlie, C. & Reynolds, D. (2000) *The international handbook of school effectiveness research* (London, Routledge).

Televantou, I., Marsh, H.W., Kyriakides, L., Nagengast, B., Fletcher, J. & Malmberg, L.-E. (2015) Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models, *School Effectiveness and School Improvement*, 26(1), 75–101.

Telhaj, S., Adnett, N., Davies, P., Hutton, D. & Coe, R. (2009) Increasing within-school competition: A case for department level performance indicators?, *Research Papers in Education*, 24(1), 45–55.

Tymms, P. & Dean, C. (2004) *Value-added in the primary school league tables: A report for the National Association of Head Teachers* (Durham, Curriculum, Evaluation Management Centre).

Visscher, A.J. (2001) Public school performance indicators: Problems and recommendations, *Studies in Educational Evaluation*, 27(3), 199–214.

Wiliam, D. (2010) Standardized testing and school accountability, *Educational Psychologist*, 45(2), 107–122.

Wilson, D. & Piebalga, A. (2008) Performance measures, ranking and parental choice: An analysis of the English school league tables, *International Public Management Journal*, 11(3), 344–366.

Appendix A

Table A1. Summary of 'Making Good Progress' performance data

Performance by time period Variables used in Analysis 3	Obs	Mean	Std. Dev.	Min	Max
2007/08 Teacher-assessed mathematics point score[1]	99,513	27.03	7.96	9	53
2008/09 Teacher-assessed mathematics point score[1]	97,044	27.31	8.06	9	53
2009/10 Teacher-assessed mathematics point score[1]	71,144	27.35	7.96	9	53
Key Stage 2 Average point score[2]	85,456	27.41	4.10	0	35.36
Key Stage 1 Average point score[3]	13,5962	15.27	3.61	3	27
Free school meals eligible	20.1%				
Gender recorded as male	52.3%				

Performance by time period and National Curriculum year

National Curriculum year	2007/2008			2008/2009			2009/2010		
	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
3	13,162	18.16	3.22	12,934	18.16	3.17	9,712	18.29	3.18
4	13,827	20.73	3.82	13,319	20.95	3.75	10,185	21.15	3.69
5	13,921	23.79	4.63	14,000	23.79	4.51	10,443	24.15	4.38

*. (Continued)

Performance by time period and National Curriculum year									
	2007/2008			2008/2009			2009/2010		
National Curriculum year	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
6	14,184	27.19	4.77	13,831	27.45	4.73	10,570	27.81	4.68
7	14,644	29.84	6.01	14,546	30.27	5.82	10,066	30.55	5.92
8	14,858	32.37	6.68	14,296	33.13	6.59	10,433	33.29	6.61
9	14,917	35.46	7.40	14,118	36.13	7.38	9,735	36.18	7.19

[1] Teacher-assessed mathematics point score data from the Making Good Progress data were recorded in sub-levels. Possible scores ranged from 9 to 53 in increments of 2. The data were approximately normally distributed with no obvious ceiling or floor effects.

[2] The Key Stage 2 average point scores were fine-grained and approximately normally distributed with the exception of a) a small floor effect for 1.2% of pupils at 15 points and b) a moderate negative skew suggesting either an under-performing tail on the distribution or a ceiling effect on higher-ability pupils.

[3] The Key Stage 1 average point scores were moderately fine-grained, with 1/3 and 1/2 points recorded as well as integers. There was a marked ceiling effect at 21 points, affecting about 8% of pupils, and a marked spike at the expected score of 15 affecting just under 16% of pupils, both serious problems with the KS1 prior attainment data.

Appendix B

Table B1. Results of a school-level multiple regression analysis of the 2014 KS2–4 Best 8 value-added measure on a number of intake characteristics

	Coefficient value	Standard error	t
Percentage[1] of pupils[2] with special educational needs (SEN)	−3.2	7.6	0.42
Percentage of pupils with English as an additional language (EAL)	74.0	2.6	30
Percentage of pupils on free school meals (FSM) or with looked-after status	−76.8	3.7	21
Total number of pupils at the end of KS4	0.014	0.006	2.1
Cohort average KS2 attainment (APS)	1.5	0.4	4.2
Coverage (percentage of pupils included in the measure)	−8.8	8.5	1.0
Percentage of female pupils	18.3	2.2	8.5

Model $R^2 = 0.35$, n = 2,990 schools, all figures to 1DP apart from total pupils, reported to 3DP.
[1] All percentages expressed as a decimal (coefficient value represents effect at 100%).
[2] All figures relate to the cohort rather than the overall school.Table B2. Results of a school-level multiple regression analysis of the 2014 KS1–2 value-added measure on a number of intake characteristics

	Coefficient value	Standard error	t
Percentage[1] of pupils[2] with special educational needs (SEN)	−1.5	0.1	13.1
Percentage of pupils with English as an additional language (EAL)	1.1	0.0	24.2
Percentage of pupils on free school meals (FSM) or with looked-after status	−0.9	0.1	17.2
Total number of pupils at the end of KS2	−0.0	0.0	13.6
Cohort average KS1 attainment (APS)	−0.2	0.0	25.0
Coverage (percentage of pupils included in the measure)	0.0	0.2	0.1
Percentage of female pupils	−0.0	0.1	0.2

Model $R^2 = 0.10$, n = 14,292 schools, all figures to 1DP
[1] All percentages expressed as a decimal (coefficient value represents effect at 100%).
[2] All figures relate to the cohort rather than the overall school.

Appendix C

Table C1. Pairwise correlations in raw scores for different cohorts in the same school at the same time for three study years (2008–2010)

Primary level Study year		1 NC year lower	2 NC years lower	3 NC years lower
2008	NC Year 6	0.64	0.56	0.50
2009		0.67	0.54	0.51
2010		0.71	0.61	0.60
2008	NC Year 5	0.70	0.61	
2009		0.68	0.65	
2010		0.74	0.67	
2008	NC Year 4	0.64		
2009		0.67		
2010		0.69		
Column mean		0.68	0.61	0.54

Secondary level Study year		1 NC year lower	2 NC years lower
2008	NC Year 9	0.86	0.83
2009		0.86	0.76
2010		0.76	0.65
2008	NC Year 8	0.83	
2009		0.88	
2010		0.86	
Column mean		0.84	0.75